2010

Desiderata for a Predictive Theory of Statistics

Bertrand Clarke

Bayesian Analysis (2010) **5**, Number 2, pp. 283–318

Desiderata for a Predictive Theory of Statistics

Bertrand Clarke*

Abstract. In many contexts the predictive validation of models or their associated prediction strategies is of greater importance than model identification which may be practically impossible. This is particularly so in fields involving complex or high dimensional data where model selection, or more generally predictor selection is the main focus of effort. This paper suggests a unified treatment for predictive analyses based on six 'desiderata'. These desiderata are an effort to clarify what criteria a good predictive theory of statistics should satisfy.

Keywords: validation, online prediction, Prequentialism, bias variance analysis, overall sensitivity, model reselection

1 Validation

The central issue in this paper is how to ensure inferences from a specific data set validate beyond a specific context. To address the validation challenges, we propose a list of six desiderata that may provide a framework for addressing a large class of statistical problems predictively. These six desiderata are not formal enough to be a paradigm or analytic framework, however, they are formal enough that their prescriptions can be meaningfully interpreted. For the sake of giving it a name, this set of desiderata is called a Coordinating Theory in the hope that it may inter-relate many of the foundational ideas in statistics. It will be seen that the approach advocated here is not purely Bayesian, but is much closer to Bayesian than it is to any other existing school of statistical thought.

The six desiderata themselves are an effort to complete the Prequential approach, see Dawid (1984) and below, to an effective alternative to the main philosophies such as Bayes, Frequentist, Conditional Likelihood and Information-theoretic. Note that use of the word "complete" is meant to convey the belief that, while many properties of the Prequential approach are known, there are statistical issues conceptually disjoint from online prediction (narrowly defined) that are important and should be integrated into Prequentialism to help make it a full prescription for statistical analysis. The benefits of proposing an ideal analysis include identifying what features of modeling are generically most important to address and identifying what features of a statistical problem are most important to be used in further study. We hope this Coordinating Theory will be well suited to complex data for which conventional modeling does not scale up well.

Briefly, the six desiderata are (i) predictive optimality subject to the Prequential principle, (ii) the use of prediction errors to update the prediction scheme, (iii) a proper

*Department of Medicine, Center for Computational Sciences and the Department of Epidemiology and Public Health, University of Miami, Miami, FL, mailto:bclarke2@med.miami.edu

 DOI:10.1214/10-BA604

treatment of all sources of variance and bias, (iv) that the complexity of the prediction procedure should be related to the difficulty of approximating the true model, (v) a complete robustness analysis, and (vi) a sanity check that the procedure behaves properly in limiting senses. Taken together these six desiderata will be termed a Coordinating Theory since they are intended to inter-relate a collection of ideas coherently.

It is easy to imagine alternative desiderata – many have argued that Savage's axioms for Bayesian inference, von Neumann's axioms for decision theory or Rostek's axioms for percentile based inference are quite satisfactory and little more is needed. Also, it is not hard to imagine revamping the set of six desiderata by using ideas from sequential hypothesis testing or data compression and transmission in place of those for prediction. However, the argument that the six desiderata presented here are more appropriate than these alternatives rests on the notion that prediction is the central optimality property to seek and that there are subsidiary criteria that we want an ideal prediction scheme to satisfy. Thus, the Coordinating Theory perspective on batch analysis in general is not that it is bad but, rather, if batch analysis succeeds, it does so by achieving good prediction, in effect satisfying the six desiderata.

The three most novel features of the six desiderata are that they introduce the concept of going outside the Bayesian paradigm to rechoose a decision problem in (ii), a comprehensive variance bias decomposition in (iii) and a complete robustness analysis in (v), distinguishing between local and global perturbations. It must be reiterated that the point of the desiderata is to provide structure for a generic problem, not to criticize existing methods per se. The criticism of methods below based on validation probably arises from the use of those methods under criteria derived from fit rather than validation. The intent here is to situate existing methods in a predictive structure to compare their performance.

It will be seen that the relationship between a Coordinating Theory and Bayes or Frequentist philosophies is intricate. From the Coordinating Theory perspective, other schools of thought such as Bayes, Frequentist, Conditional Likelihood, and Information-theoretic are techniques by which to construct predictors with certain properties. Coordinating Theory would be a way to compare the predictors, on the basis of the six desiderata. The setting of desideratum (i) is decision-theoretic and hence more Bayesian than Frequentist (since the class of Bayes actions is complete). Desideratum (ii) is, at root, a Frequentist criterion meant to quantify bias and variance over repeated sampling. Desideratum (iii) recognizes that Bayes is optimal in a given decision problem but that choosing which decision problem to solve to get a good predictor is usually not Bayesian. Desideratum (iv) is neither Bayes nor Frequentist; it is imported from the minimum description length philosophy of statistics. Desideratum (v) invokes a sort of objective Bayes robustness. Desideratum (vi) is again a Frequentist property from repeated sampling. Taken together, the Desiderata provide much more structure for inference than the usual Machine Learning view which rests on predictive performance almost exclusively.

The structure of this paper is as follows. In the next Section the conventional modeling framework is discussed and evaluated and the Prequential approach is described

as the basis of an alternative. In Section 3, the six desiderata of Coordinating Theory are stated and explained. Section 4 then discusses the relationship of the Coordinating Theory to existing schools of thought in Statistics, focusing on Bayes and Frequentist thinking. In Section 5, several aspects of the first four desiderata are explored in a computed example. Section 6 discusses several of the implications of the perspective developed here.

2 Conventional Modeling vs Prequentialism

The conventional model for inference rests on using outcomes of random variables such as $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to select from a class of objects say $\mathcal{P}$. Often $\mathcal{P}$ is a parametric family, but may be a nonparametric class of densities, distributions, or functions or a class of decisions, among other possibilities. A probability measure P_X can be induced on the image of X in which case the range measure space is written $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$; it is the "visible" part of the random variable. That is, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ is the part of a phenomenon that we model; the underlying triple, $(\Omega, \mathcal{F}, P)$, is essentially masked from us and not modeled. The idea is that the macroscopic world of real data that we analyze arises from an unseen world and we use our data analysis and $\mathcal{P}$ in a effort to uncover P. From this point, mainstream statistical analysis roughly separates into two philosophies, Bayes and Frequentist.

2.1 Limitations of the Conventional Model

Conventional models, while amazingly successful, can be criticized from two directions. The first direction is that it is not at all clear how well a given conventional model scales up to high dimensional and complex data. Indeed, it is not necessarily clear that the idea of a true model is even very useful in settings with complex and high dimensional data. Wrong but simple models that are good approximations to a complex true model may give better results than the true model for realistic sample sizes and certain collections of covariates, see Yang (1997) and Wainwright (2006) for instance.

In the Bayesian context, the problem with scaling up to high dimensions is seen in the phenomenon of dilution – a sufficient number of wrong but good models may split the posterior probability so finely that a genuinely poor model may appear better, see Chipman and McCulloch (1992). An extension of this idea is that the posterior may in fact converge to zero if the model list is permitted to grow too quickly and includes enough good models that they all dilute. Steps can be taken to correct this, see George (2001), but they rest on 'uniformizing' the prior over models on neighborhoods severely limiting the choice of prior. Pericchi (2005) reviews objective priors for Bayes factors in model selection. He proposes that lack of robustness in prior selection can some times be overcome by techniques such as intrinsic priors or expected posterior priors. When such techniques give similar results, i.e., we have robustness, the inferences may be more reliable. Nevertheless, prior selection and posterior exploration in high dimensional contexts remains hard.

Current efforts to scale up linear regression models to high dimensions via shrinkage criteria include SCAD (Fan and Li (2001)), Elastic Net (Zou and Hastie (2005)), Adaptive LASSO (Zou (2006)), and Adaptive COSSO (Storlie et al. (To appear)) among others. This class of techniques derives its justification from the oracle property, see Fan and Li (2001) that effectively requires p not be too large relative to n. In practice, it is unclear in what this growth rate means, but for the Adaptive elastic net, $p = n^\gamma$ for $\gamma < 1$, see Zou and Zhang (2009) and for the SCAD penalty Huang and Xie (2007) require $p = o(\sqrt{n})$. Commonly this is violated by -omics data where $p \approx 20,000$ and $n \approx 50$. These methods may reduce to Hodges' super-efficiency, see Leeb and Pötscher (2001). Another open question is the assignment of SEs to all parameter estimates in these methods. Also, it is unclear how well they perform relative to forward, backward, and stepwise selection, and they can give poor results with dependent covariates.

There are other methods such as clustering and data summarization to achieve variable selection or dimension reduction. While these methods are helpful and often give improved performance, most have limitations. One of the most important recent techniques is called sure independence screening (SIS), see Fan and Lv (2008). While SIS does scale up to the case that p is exponential as a function of n, it rests on correlation and so is essentially a linear, marginal procedure subject to the usual problems of such methods. (For instance, if $(Y, X_1, X_2) = (3,3,1), (4,4,1), (5,5,1), (3,8,2), (4,9,2)$ and $(5,10,2)$ both X_1 and X_2 have correlation zero with Y even though Y clearly depends strongly on both.) It is not clear yet how well this method performs on complex data. For instance, Webster et al. (2009) used SIS and permutation methods to reduce one million dimensional SNP data to around 250,000 dimensions. Even so, shrinkage methods will routinely break down in such cases.

Classical and Bayesian nonparametrics are also efforts to deal with complex and high dimensional data. The usual problem with classical nonparametrics is that it suffers the Curse of Dimensionality. There are regression techniques such as neural networks and projection pursuit that evade the Curse, see Barron (1993) and Zhao and Atkeson (1993). However, neural networks are notoriously unstable (even with regularization) and projection pursuit does not seem to have been investigated as well as it deserves. Bayes nonparametrics remains promising, but to a large extent is still under development. Gaussian process priors and Dirichlet process priors have been the main choice for several years. More recently, numerous extensions and generalizations have been proposed. See, for instance, the recent contribution of Kim et al. (2009) and the summary of Lijoi and Prünster (2009) and the references therein. Overall, prior selection remains an open question.

The second direction of criticism of conventional models is that there is a strong tendency for them to fail to validate for complex data. Ransohoff (2004) and Ransohoff (2005) document many examples from the cancer literature where the results of studies failed to stand up under repeated testing, despite the conclusions having seemed persuasive. Ransohoff (2004) identifies overfitting as the most obvious shortcoming of otherwise plausible analyses and Ransohoff (2005) states that bias is such a pervasive and hard-to-detect problem that "results are guilty of bias until proven innocent". In part, Ransohoff's examples are the logical outcome of failing to account adequately for

all the uncertainties in modeling, see Draper (1995), which are especially severe in the '-omics' fields that Ransohoff surveys. Overall, Ransohoff's criticisms of existing approaches lead him to emphasize extensive validation of results before they should be considered established. It would be naive to think that it is only in the -omics fields that validation often fails.

Overall, inference techniques derived under the conventional model are well suited to models with a relatively small number of coefficients in a relatively simple model with a relatively simple data set and decent sample size i.e., when a sparse model is "true" or at least best and not too hard to find. In some of these cases, the population for which a sparse model exists will be somewhat narrow or artificial so that actual use is open to question. Moreover, in many contexts, it is not reasonable to assume sparsity. The typical case encountered in complex and high dimensional data types, such as in the -omics world, is that the observed phenomenon is the result of many small contributions.

Taken together, these two points – difficulty of finding reliable methods in complex data settings and ensuring that they work well in reality – suggests there may be room for a reformulation of the statistical paradigm, away from the conventional model and, to an extent, away from both the orthodox Bayesian and Frequentist philosophies.

2.2 The Prequential Approach

An alternative to the conventional model is provided by the Prequential approach, see Dawid (1984). The Prequential Principle comes in three forms, weak, strong and super-strong. The weak prequential principle – that is assumed unless otherwise specified – requires that any criterion of agreement between a forecaster and a data generator (DG) should depend only on the actual observed sequences of forecasts and outcomes, and not on any of the strategies which might have been used to produce the forecasts or outcomes. The strong and super-strong prequential principles add conditions to the weak prequential principle, see Dawid and Vovk (1999). One of the undernoticed implications of the Prequential principle is that it can be interpreted to be "strictly forward" in the sense that retrodiction is ruled out. Consider one-day-ahead predictions of rain over n days. Then, using a construct such as the probability assigned to rain for 14 July on 13 July given that it did rain on July 15 would be disallowed, cf. van Erven et al. (2008). Note that there is no prohibition on using a function of data that estimates this probability.

One of the benefits of the Prequential approach is that the centrality of comparing forecasts and outcomes emphasizes the role of validation. Indeed, the Prequential principle only requires forecasts and outcomes to be compared, reducing the importance of models. This is partially because good prediction is a more general task than model identification but also because the DG is ruled out as a factor in evaluating forecasts.

The Prequential approach is motivated in part by de Finetti (1937), see also Kyburg and Smokler (1980). Essentially, de Finetti uses a subjective probability framework to issue conditional probabilities of observable events given previous outcomes instead of point predictions. In de Finetti (1937) he writes: "l'observation peut seulement nous

donner des renseignements qui sont susceptibles d'influencer notre opinion... elle signifie qu' à la probabilité d'un fait subordonné à ces renseignements – probabilité bien distincte de celle du même fait non subordonné à d'autres – nous pouvons attribuer effectivement une valeur différente," (p. 63, Chap. VI). Loosely: 'observation can neither confirm nor refute an opinion which is neither true, nor false. Observation can only give us information which is likely to influence our opinion.... this means that the probability conditional on extra information – which is distinct from the unconditional probability – can have a very different value from it.' de Finetti (1937) focusses on the simplest cases that do not have explanatory variables.

By contrast, Dawid (2004) puts great emphasis on real-world testing (Sec. 6). Indeed, Dawid (2004) writes that his approach makes no metaphysical assumptions about probability, causality, or determinism although it does 'support a straightforward approach to building, testing, using and interpreting probabilistic theories of the world.' Dawid does not require there be a probabilistic data generator or that the task of inference be to learn what it is.

One of the criticisms of the Prequential approach is it provides little guidance for analyzing a batch of data that have been collected. With batch data, often the order of collection has been lost and if independence is assumed, many would argue any ordering is artificial. The obvious answer to this is to 'batchify' the Prequential approach by choosing a number of random permutations of the data, do the analysis sequentially for each, and then average over the results to make predictions and quantify aggregate behavior like the sequence of predictive errors. This sort of procedure is used in an example in Section 4.

Separate from averaging over permutations, there are three reasons why looking at batch data sequentially is useful. First, looking for good predictors is a more general problem than finding a good model and so is a weaker criterion that has a better chance of being achievable. Roughly, every model has associated predictors, but it is not clear that every predictor corresponds to a model. Searching more broadly for predictors may result in obtaining better predictive properties than a model would have. The cost would be not having a model, but many of the interpretive properties one can derive from a model can also be derived from a predictor. Often, it will be possible to identify a model having a natural predictor that approximates a good predictor and gives satisfactory predictions, albeit not as good as the predictor itself.

Second, there are insights into the data that emerge from looking at it sequentially that are not likely to be found from a batch analysis approach. First, it is only a sequential approach that reveals some predictors routinely outperform Bayes predictors pre-asymptotically, see Wong and Clarke (2004). Likewise, Clarke (2003) has shown that Bayes predictors are often outperformed, again in small samples, when the model list does not contain the true model; i.e., in the presence of bias. These examples are related to the fact that prediction and Bayesian analysis are subtly different cf. van Erven et al. (2008). Roughly, by focussing attention on the support of the prior, the Bayesian works with a "closed mind" and can never discover that a model list is wrong without having a clear alternative, see Dawid (1982) who argues that falsification of

hypotheses (in the sense of Popper) must be added to Bayesian analysis.

Third, one of the strengths of the predictive approach to validation for batch data is that it helps avoid just fitting noise. That is, because Prequential techniques develop models sequentially, they may be less prone to over- and under-fit because extraneous terms are eventually ruled out and missing terms may be found as n increases, addressing Ransohoff (2005). In addition, in Prequential modeling validation is inherent in the procedure whereas in batch modeling only fitting is inherent in the procedure, addressing Ransohoff (2004).

To see this third point in more detail, consider comparing sequential prediction to cross-validation (CV) in a regression problem. Recall that residuals are a natural assessment of fit more than validation. However, residuals can be converted into an assessment of validation by internal prediction. So, consider a fixed sample of size n. Suppose k_1 data points are used to fit a model, k_2 data points are used for validation, and $k_3 = n - k_1 - k_2$ data points are left over. Then, to do internal validation, values of k_1 and k_2 must be chosen. It makes sense to choose more than one pair (k_1, k_2) because otherwise inferences may depend on the way the data are split. So, the question is which pairs to choose. In K-fold CV, the data points are given a fixed order and partitioned into K disjoint subsets. Each subset is held out in turn for validating a model formed using the other $K-1$ subsets. Thus, $k_1 = (K-1)n/K$, $k_2 = n/K$, and $k_3 = 0$ and K different partitions of the data are used. By contrast, in Prequential validation, i.e., online prediction, $k_1 = 0, 1, 2...$, $k_2 = k_1 + 1$ and $k_3 = n - k_2$ and $n-1$ different partitions of the data are used. Thus as n increases, the raw number of tests of a predictor is n i.e., increases. Leave-k-out CV has $\mathcal{O}(n)$ tests as well and when $k = 1$ there are exactly n of them as well. However, whether leave-1-out CV or sequential prediction is better is not clear. Nevertheless, Prequential validation in batch data is just another instance of internal validation. The reason to choose Prequential validation over cross-validation is that letting the size of the training set increase is a proxy for permitting models of increasing complexity as data accumulate.

3 Six Desiderata for Predictive Analysis

Four typical inference problems in statistics are classification, model identification, decision making and prediction. It is easy to see that classification and model identification can be expressed in terms of prediction: A good classifier or model based on a sample of size n should perform well on future outcomes compared to other classifiers or models based on the same data. Also, decision making can be evaluated predictively: Rather than using a decision rule to compare two hypotheses, for instance, convert both hypotheses into predictors and evaluate their performance. The hypothesis with the lower prediction error is selected as true. (For $\mathcal{H}_1 : \theta \in R_1$ vs. $\mathcal{H}_2 : \theta \in R_2$, given data $\mathcal{D}_n = \{(Y_i, \mathbf{X}_i) : i = 1, ..., n\}$ where the $\mathbf{X}_i$'s are covariates, one can convert hypotheses to predictors $\hat{F}_{j,i}$ for $j = 1, 2$ by writing $\hat{F}_{j,i} = E(Y_{i+1}\mathbf{1}_{R_j}(\Theta)|\mathcal{D}_i)$, i.e., mixing over the sets in the hypotheses.) More generally, the decision with lower predictive error is preferred. Thus, it may be reasonable to use a generic prediction problem as the central

setting of statistics instead of measure-theoretic probability as for conventional models.

So, suppose the task is to predict the next outcome Y_{n+1} using $\mathbf{X}_{n+1}$ and to determine the predictor $\hat{F}_n$ given a stream of data $\mathcal{D}_n$. Here, for each time step i, a value of $\mathbf{X} = (X_1, ..., X_p)$ is a set of explanatory variables for Y_i. Thus, we write the prediction for time $n+1$ as $\hat{Y}_{n+1} = \hat{F}_n(\mathbf{X}_{n+1})$. Now, we can state the six desiderata.

Desideratum #1: Evaluate a predictive scheme by its online cumulative predictive error subject to the Prequential principle.

The intuition behind this desideratum is that we must have some way to evaluate how well a predictive scheme (our modeling) has performed relative to reality and that the evaluation must be fair. The fairness is imposed by the Prequential principle: When comparing two predictors it is fair if (i) we only look at how each predictor makes use of the data and (ii) no predictor uses information about the model that is unavailable to other predictors.

There are two ways this desideratum can be interpreted. The first is for stochastic data, the second is for data types that cannot plausibly be regarded as the outcome of random sampling. The key difference is whether we are willing to assume the data arise from a probability model or whether we are merely invoking a probability model for use in an analysis. For instance, the sequence of letters in a novel cannot be realistically modeled as a series of outcomes of a letter-valued random variable however if we wanted to transmit the novel letter by letter we might use a Shannon code which comes from treating the letters as if they were outcomes of a random variable.

Since it is more familiar, we begin with the stochastic case. For stochastic data, the cumulative prediction error, or empirical risk, can be written as

$$CPE(n) = \frac{1}{n} \sum_{i=1}^{n} L(\hat{Y}_i, Y_i), \tag{1}$$

for a loss function L, where it is understood that the data points are ordered so that only points up to stage $i-1$ are used to predict Y_i. One instantiation of this is to use the mean squared predictive error (MSPE) in which L is taken to be squared error loss. The CPE is calculated at each time step i and represents the sum of the sequence of residuals from the use of a sequence of predictors $\hat{F}_i$ for $i = 1, ..., n$ from a specified predictive process. Since $e_i = Y_i - \hat{Y}_i$ is the residual from a predictor, e_i is often called a predictual. More generally, the error $L(\hat{Y}_i, Y_i)$ is also called a predictual. Note that using the CPE is equivalent to online prediction: (1) can be found from the one step errors from strictly online prediction and the one step errors from online prediction can be recovered from $CPE(1)$,..., $CPE(n)$.

If $CPE(n)$ converges to a limit $E(CPE)$, then $E(CPE)$ may be minimized in principle. However, this is disallowed by the Prequential principle because it depends on the model under which E is taken. Minimization of $E(CPE(n))$ for any fixed n has the same problem. However, $CPE(n)$ can be minimized. While this may be reasonable in

some cases, if the class of predictors over which the optimization is done is large enough to be realistic, $\arg\min CPE(n)$ will degenerate to a function that matches the data $\mathcal{D}_n$ perfectly but has poor generalization error. Thus, we only use the CPE as in 1 to evaluate predictors. It is easy to imagine variations on Desideratum #1. For instance, a median predictive error such as $\text{Median}_{i=1,\ldots,n} L(\hat{Y}_i, Y_i)$ could be used in place of (1). Other operations on the predictuals are possible.

Another feature of Desideratum #1 is its use with retrodiction. When the Prequential principle is invoked, retrodiction is sometimes taken to be ruled out on the grounds that predictions must always be forward. However, CV and online prediction are internal prediction criteria for batch data. So, consider the setting that M predictors are available and suppose each is derived from a specific model for the data. Then, if the data is IID or stationary more generally, it may make sense to ignore the ordering up to and including the $i-1$ time step when making a prediction for time i. Thus, the sequentiality of the prediction is maintained for present-time use but all the accumulated data can be used in a less restricted fashion. In this case, one could use K-fold CV with all the data at time $i-1$ to select one of the M models to make a prediction at time i.

Now suppose we have a nonstochastic data type that cannot plausibly be regarded as the outcome of random sampling. An example would be a long vector $(x_1, ..., x_p)$ with regularities. Since neither a mean nor a variance is a reasonable summary, a variant on (1) would be better.

Start by regarding online prediction as a sequential game between Nature, N, and a Forecaster, F, permitting the Forecaster access to a collection $\mathcal{E}$ of experts indexed by θ where $\theta \in \mathcal{E}$, see Shtarkov (1988), Haussler and Barron (1992), and Haussler et al. (1998). Each round of the game is organized by a Referee. The order of play is that the Referee obtains the opinions of the experts, tells F and then receives F's density from which predictions will be made. Then, the Referee receives N's choice of outcome x, and calculates how much F must pay N or the reverse.

Suppose each round of the game uses a log scoring rule. Then, to start the game, each expert θ announces a density $p_\theta(\cdot)$ for x_1. The Forecaster sees this and tries to match the performance of the best expert in $\mathcal{E}$ by choosing the density $q(\cdot)$ from which to make predictions for x_1. Then, N chooses the actual value of x_1 arbitrarily. The Referee then makes F pay $\ln 1/q(x_1)$ to N. So, the question is how F should choose q. Note that an expert would pay $\ln 1/p_\theta(x_1)$ and the best expert would pay $\min_\theta \ln 1/p_\theta(x_1)$. Thus, F might try to minimize the amount lost beyond what the best expert would lose.

Thus, F should choose q to minimize the difference

$$\ln 1/q(x_1) - \inf_\theta \ln 1/p_\theta(x_1) = \sup_\theta \ln \frac{p_\theta(x_1)}{q(x_1)}. \tag{2}$$

Expression (2) is called the regret. Clearly, the worst N could do to F would be to choose x_1 to maximize (2). So, F might be led to choose q to achieve the minimax value

$$\inf_{q \in \mathcal{P}} \left(\sup_{x,\theta} \log \frac{p_\theta(x)}{q(x)} \right). \tag{3}$$

Shtarkov (1988) proved that the optimal q is

$$q_{opt}(x) = \arg_q \left[\inf_{q \in \mathcal{P}} \left(\sup_{x,\theta} \log \frac{p_\theta(x)}{q(x)} \right) \right] = \frac{p(x_1|\hat{\theta})}{\int p(x|\hat{\theta}(x))dx} \tag{4}$$

the normalized, maximized likelihood. When (4) exists, the value of (3) is the log of the normalizing constant in q_{opt} and when x is a vector of length n, say x is replaced by the vector $(x_1, ..., x_n) = x^n$, it has an asymptotic form (in n) given by Rissanen (1996), eq. 6. The Shtarkov solution q_{opt} is an average of the models over the sample space, but the n-th solution (for x^n) is not the marginal from the $n+1$ solution (for x^{n+1}). Xie and Barron (2000) give a complete minimax analysis of this case for discrete xs.

In the Bayesian version of this game, the experts are taken as subjectively weighted by a prior $w(\theta)$ based on their reliability. So, (4) becomes

$$\arg_q \left[\inf_{q \in \mathcal{P}} \left(\sup_{x,\theta} \log \frac{w(\theta)p_\theta(x)}{q(x)} \right) \right] = \frac{w(\tilde{\theta})p(x|\tilde{\theta})}{\int w(\tilde{\theta})p(x|\tilde{\theta}(x))dx}, \tag{5}$$

a variant on Shtarkov (1988) and Rissanen (1996), where $\tilde{\theta}$ is the posterior mode rather than the MLE. The asymptotics for (5) when x is replaced by x^n are in Clarke (2007).

To put this in the setting of Desideratum #1, imagine F plays n rounds of the Bayesian game (5); this will reduce to the worst case individual sequence or Frequentist game if $w \equiv 1$. So, use $x = x^n$ but examine the sequence of n univariate games for each x_i, $i = 1, ..., n$. Omitting subscripts on the qs and sequentializing the Bayesian version of the Shtarkov game means finding

$$\arg \left[\inf_{q \in \mathcal{P}} \left(\sup_{x_i,\theta} \ln \frac{w(\theta|x^{i-1})p_\theta(x_i|x^{i-1})}{q(x_i|x^{i-1})} \right) \right]. \tag{6}$$

So, set

$$\tilde{\theta}_i = \arg \max_\theta w(\theta|x^{i-1})p_\theta(x_i|x^{i-1})$$

to see that the optimal density for the i-th round is now

$$q_{opt,i}(x_i|x^{i-1}) = \frac{w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})}{\int w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})dx_i}. \tag{7}$$

The value of (6) is now

$$\ln \int w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})dx_i.$$

The cumulative regret is now the CPE for the non-stochastic case, parallel to (1) for the stochastic case. For the Shtarkov predictor, write the CPE as

$$CRegret(n, x^n) = \sum_{i=1}^n \ln \frac{w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}}(x_i|x^{i-1})}{q_{opt,i}(x_i|x^{i-1})} = \sum_{i=1}^n \ln \int w(\tilde{\theta}_i|x^{i-1})p_{\tilde{\theta}_i}(x_i|x^{i-1})dx_i.$$

Note that predictive performance requires specification of a class of predictors and a criterion for good prediction both at time n, but there is no necessity that the class or criterion must be constant as a function of n. In some cases, this means that the Prequential principle can lead to unexpected results. Wong and Clarke (2004) developed "mongrel" predictors based on re-selection of the error criterion used to form them and showed they outperformed Bayes methods in small samples in a Prequential sense in the stochastic case. This verifies that while Bayes methods can be asymptotically optimal they needn't be finite sample optimal; this holds for the Shtarkov setting as well. In essence, standard Bayes methods can be outperformed predictively when the decision problem that a predictor solves at time $n+1$ is allowed to be different from the decision problem it solves at time n. That is, the criterion and action space e.g., the model list, used to choose a predictor is updated from time n to time $n+1$, not just the the predictor itself; this happens implicitly in van Erven et al. (2008). The interpretation of this kind of extensive updating is that over time, we refine our idea of what problem to solve to find good predictors.

Desideratum #2: Use Prediction Errors to Update the Prediction Problem.

In the sequential prediction problem, a prediction must be made at each time step based on the accumulated data. So, imagine fitting a linear regression model using a data set and making a prediction from it. When the next data point is revealed, there are two possibilities. First, the prediction is good in the sense of (1) say and the model is validated. Then, it is enough to update the parameter estimates with the new data point and make another prediction. Second, the prediction is deemed inadequate: The new data point gives an excessive value for (1), is outside the 95% prediction interval, leads to too many predictions that are systematically higher than their observed y_i's (although $L(y_i, \hat{y}_i)$ is small), or leads to some other undesirable feature. In this case, following Dawid (1992), we may be led to sequential reselection of the prediction problem before estimating any parameters. Thus, Desideratum #2 is little more than residual analysis done predictively.

To see this more formally, suppose reality chooses F to be true and we first choose a space of functions, $\mathcal{F}$, to help form $\hat{F}$. Then, we might set up a decision problem for predicting the next outcome as follows. Fix an initial list $\mathcal{M}_0 \subset \mathcal{F}$ with M_0 items, where $\mathcal{M}_0 = \{f_{0,1}, f_{0,2}, ..., f_{0,M_0}\}$ is the action space $\mathcal{A}_0$ and the loss L is squared error. Put a prior W on the list $\mathcal{M}_0$, and assign priors for any parameters within each model $f_{0,k}$. Since the Bayes model average (BMA) minimizes the posterior risk, it is the Bayes action from which to make predictions at stage n.

Now, suppose that as an experimenter accumulates data, the CPE grows so large that for $n > n_0$ the adequacy of the modeling strategy is doubtful. Perhaps the list $\mathcal{M}_0$ was wrong (too large, too small, poorly located etc.) thereby giving a poor action space $\mathcal{A}_0$ with high CPE. Perhaps the loss function or the priors, or even $\mathcal{F}$ is found to be wrong. Then, the experimenter might reformulate the prediction problem with new choices of the model list, optimality criterion, or action space. After problem reselection, the CPE based on all the accumulated data would still be used.

Repeating these reselections, in principle for each time step, one version of the general problem may be stated as follows. For each n, find $\mathcal{M}_n \subset \mathcal{F}$ with $\mathcal{M}_n = \{f_{n,1}, ..., f_{n,M_n}\}$, finitely many finitely parametrized models to be used at time n, and find a criterion under which to choose $\hat{F}_n$ as a model average from $\mathcal{M}_n$, with weights determined by the data up to time n.

An instance of reselecting the loss function would be the following. First, distinguish between the loss function used to calculate the CPE and the loss function used to construct the predictor. So, suppose predictions will be evaluated under squared error loss. Then, in small samples with data exhibiting high variability, the median might be a better predictor than the mean. That is, for small samples, the median – which corresponds to using L^1 loss – may give better predictive performance initially even though it is L^2 loss we want to minimize. Past a certain value of n, of course, we would revert to L^2 loss, and hence the mean. This happens for n small enough that the higher efficiency of the least squares estimator does not overcome the greater stability of the median. Also, this phenomenon has been observed in van Erven et al. (2008) who switch from selecting a model using AIC to selecting a model using BIC. Another instance would be using an initial loss function for which stacking was the optimal action until a model list with low enough bias had been found so that BMA (under L^2) would be appropriate. In principle, there is no prohibition on changing the loss function used in the CPE. An instance of this would be using L^p loss and letting p increase with n. This would have the effect of penalizing errors more and more heavily as n increased.

Note that updates of the prediction problem can have new optimal actions that might or might not be Bayes. Indeed, it is likely that this reselection of the prediction problem to solve cannot be done in a Bayesian fashion. Because the prediction problem is reformulated from time to time, the prediction scheme effectively searches over prediction problems to find the right one to solve to obtain good prediction. That is, the early prediction problems and their solutions are estimates of the correct prediction problem and its solution.

Desideratum #3: Generate a Unified Bias-Variance Analysis.

The intuition is that the nature of the uncertainty associated to each input to a predictive scheme should be examined. Variance and bias are two established assessments of uncertainty, but, as seen below, data compression formulations may also be relevant in the non-stochastic data case. It is hard to state clearly in generality how Desideratum #3 can be implemented so consider the following five examples.

Begin with a very simple predictive system using one model that is known up to finitely many parameters denoted $\beta = (\beta_1, ..., \beta_p)$. The only inference problem is to estimate β. Provided there is a value β_T that makes the model correct, the predictor based on the plug in estimate $\hat{\beta}$ should give better and better predictions as it converges to β_T. That is, parameter consistency gives predictive optimality asymptotically. In

this case, there is only one bias-variance decomposition based on the mean squared error

$$MSE(\beta_T, \hat{\beta}) = \sum_{j=1}^{p} \mathsf{bias}(\hat{\beta}_j, \beta_{j,T})^2 + \sum_{j=1}^{p} \mathsf{Var}(\hat{\beta}_j), \tag{8}$$

in which $\mathrm{bias}(\hat{\beta}_j, \beta_{j,T}) = E\hat{\beta}_j - \beta_{j,T}$ and the expectations and variance are taken in the distribution indexed by β_T.

As an extension of (8) to function estimation, Domingos (2000) established a decomposition of the prediction risk into three terms, valid for a class of loss functions that includes zero-one and squared error. Two of the terms are recognizable as variance and bias while the third represents the irreducible noise of the data generator. This analysis rests on the regression model $Y = F(X) + \epsilon$ in which F is estimated by $\hat{F}(\cdot) = \hat{F}(\cdot; \mathcal{D}_n)$ where $\mathcal{D} = \mathcal{D}_n$ is a sample of size n. The Domingos (2000) analysis would apply to a nonparametric predictor that had no inputs. This might correspond to, say, kernel regression in which the tuning parameter were chosen as a fixed value independent of the data so its value could be treated as a robustness issue under Desideratum #5.

The Domingos (2000) decomposition posits a predictor $\hat{Y}(\mathbf{x}) = \hat{F}(\mathbf{x}; \mathcal{D})$ evaluated a fixed $\mathbf{x}$. Assign a 'main prediction' $y_m(\mathbf{x})$ to $\hat{Y}(\mathbf{x})$ by defining

$$y_m(\mathbf{x}) = \arg\min_{y'} E_{\mathcal{D}} L(\hat{Y}(\mathbf{x}), y'), \tag{9}$$

in which the expectation is taken with respect to the sampling distribution for $\mathcal{D}$ and L is a loss function. If L is squared error loss, then $y_m(\mathbf{x}) = E_{\mathcal{D}}\hat{Y}(\mathbf{x})$. Now, it makes sense to define the variance of the predictor $\hat{Y}(\mathbf{x}) = \hat{F}(\mathbf{x})$ to be

$$\mathsf{Var}(\hat{Y}(\mathbf{x})) = E_{\mathcal{D}} L(y_m(\mathbf{x}), \hat{Y}(\mathbf{x})). \tag{10}$$

Note that (9) and (10) are Case I in James and Hastie (1997) who used absolute error and zero-one loss. This is consistent with Heskes (1998) who used relative entropy loss.

Turning to the random variable $Y(\mathbf{x})$ we want to predict, let $y^* = y^*(\mathbf{x})$ be its 'ideal predictor' defined by

$$y^*(\mathbf{x}) = \arg\min_{\hat{y}} E_{noise} L(Y(\mathbf{x}), \hat{y}), \tag{11}$$

in which the expectation is in the distribution of the noise term ϵ. That is, $y_m(\mathbf{x})$ would be the best predictor under L if we knew $F(\mathbf{x})$. If L were squared error, $y^*(\mathbf{x}) = E_{noise} Y(\mathbf{x})$.

The problem is that, in general, $y_m(\mathbf{x}) \neq y^*(\mathbf{x})$. So, Domingos (2000) defines

$$\mathsf{Bias}(\mathbf{x}) = L(y^*, y_m). \tag{12}$$

Now, given (10) and (12), the proof of Theorem 1 in Domingos (2000) establishes that an analog to (8) for the function estimation case is

$$\begin{aligned} E_{\mathcal{D},noise} L(Y(\mathbf{x}), \hat{Y}(\mathbf{x})) &= E_{noise} L(Y(\mathbf{x}), y^*(\mathbf{x})) + L(y^*(\mathbf{x}), y_m(\mathbf{x})) \\ &+ E_{\mathcal{D}} L(y_m(\mathbf{x}), \hat{Y}(\mathbf{x})) \\ &= N(\mathbf{x}) + \mathsf{Bias}(\mathbf{x}) + \mathsf{Var}(\hat{Y}(\mathbf{x})), \end{aligned} \tag{13}$$

for a variety of loss functions. The term $N(\mathbf{x})$ in (13) is the inherent variability of the data generator regardless of any $\hat{Y}$. It is effectively a variance for Y, so (13) has two variance terms and one bias term.

Note that (8) and (13) each involve one variance around the true value of the parameter and around the main prediction, respectively. Likewise, (8) and (13) each involve one bias away from the true value of the parameter and away from the ideal predictor, respectively. That is, in these two examples, there was only a single unknown input to the prediction scheme, the unknown parameter or the unknown function value, and so only a single variance and bias to examine. The term unified bias-variance analysis refers to using several versions of (8) or (13) to analyze all inputs to the predictor. First, there are three kinds of input, apart from the data. Some inputs are knowable in the sense that they can be estimated from the data. Some inputs are not estimable – a prior density would be the paradigm case. Some inputs are taken as known, e.g., it might be known that the best way to analyze a signal is in its Fourier basis. The third class of input will be examined in Desideratum 5 since if an input is known it doesn't have a meaningful bias or variance apart from zero.

Each input that is estimable from the data, such as β or F in the last two cases, requires a variance-bias analysis. By contrast, non-estimable inputs require a variance analysis only since they must be chosen and this will have uncertainty but they do not have a true value that can be used to form a notion of bias.

To see how a unified bias-variance analysis can be done for more than one estimable input, consider a simple case of Frequentist model averaging. Suppose $Y = F(\mathbf{x}) + \epsilon$ in which $F \in \mathcal{F}$, a space of functions, and let the class of terms that can be used in a linear model predictor for F be $\mathcal{E}$. So, the full list of models is $\mathcal{M} = \mathcal{M}(\mathcal{E})$ with cardinality $\mathsf{card}(\mathcal{M}) = m$. A typical element of $\mathcal{M}$ is $f(\mathbf{x}|\beta) = \mathbf{x}_f^T \beta_f$ where $\mathbf{x}_f$ is the vector of terms from $\mathcal{E}$ defining f and β_f is the corresponding vector of coefficients.

Given data $\mathcal{D}$, the stacking average, see Wolpert (1992), is

$$\hat{F}_{\text{stack}}(\mathbf{x}_{n+1}) = \sum_{j=1}^{m_n} \hat{w}_{n,j} \hat{f}_j(\mathbf{x}_{n+1}), \tag{14}$$

where $\hat{f}_j(\mathbf{x}) = f_j(\mathbf{x}|\hat{\beta}_f)$ for some estimator and $m_n = \mathsf{card}\mathcal{M}_n$ is the cardinality of the set $\mathcal{M}_n$ of models used to form the average for time $n+1$. The weights $w_{n,j}$ are found by a CV-type procedure. Let $f_j^{(-u)}(\mathbf{x})$ be the prediction at any $\mathbf{x}$ using model j, as estimated from training data with the u-th observation removed. Then the estimated weight vector $\hat{w}_n = (\hat{w}_{n,1}, ..., \hat{w}_{n,m_n})$ solves

$$\hat{w}_n = \text{argmin}_w \sum_{u=1}^{m_n} \left[y_u - \sum_{j=1}^{m_n} w_j \hat{f}_j^{(-u)}(\mathbf{x}_u) \right]^2. \tag{15}$$

This puts low weight on models that have poor leave-one-out accuracy; five or ten-fold cross-validation would be better in practice but is not needed for the present argument.

The class over which w_n is optimized can make a difference; natural choices include no constraints, all $w_{n,j} \geq 0$, and $\sum_j w_{n,j} = 1$. (Equation (15) does not satisfy the first desideratum; however, it can be modified so that the only data used to estimate $\hat{f}_j^{(-u)}$ precedes u.)

Now, suppose that under some optimality criterion, the best predictor of Y_{n+1} using a linear combination of models in $\mathcal{M}$ based on $\mathcal{D}$ can be defined. For instance, parallel to (11), this might be given by the list, weights, and parameter values

$$\begin{aligned}(\mathcal{M}_T, \quad (\quad & w_1^T, ..., w_{m_T}^T), (\beta_1^T, ..., \beta_{m_T}^T)) \\ = \quad & \arg \min_{A, w_1, ... w_{\mathsf{card}(A)}, \beta_1, ..., \beta_{\mathsf{card}(A)}} E\|F(\mathbf{x}) - \sum_{j \in A} w_j f_j(\mathbf{x}|\beta_f)\|, \end{aligned} \tag{16}$$

where $A \subset \mathcal{M}$, $m_T = \#\mathcal{M}_T$, $\|\cdot\|$ is the norm based on the Euclidean inner product, and the β_fs are to be estimated, for instance by MLEs from a sample of size n.

To obtain a unified bias-variance analysis for the present case, note there are three estimable inputs, namely, the parameters θ_f, the weights $\hat{w}_n$, and the model list $\mathcal{M}_n$. Using (16) in place of (11), an analog to (8) can be constructed for each of the estimable inputs. First, the "true" model list $\mathcal{M}_T$ can be regarded as a vector $\mathbf{e} = (e_1, ..., e_{\mathsf{card}(\mathcal{M})})$ of zeros and ones and the estimated model list can be regarded as $\mathbf{a} = (a_1, ..., a_{\mathsf{card}(\mathcal{M})})$ where $a_j = 1$ if model j is in $\mathcal{M}_n$ and zero otherwise. In $\mathbf{a}$, exactly m_n of the entries are ones and $a_j = a_j(\mathcal{D})$ where j indexes the j-th f in $\mathcal{M}$. Now, parallel to (8), the bias-variance expression for the model list is

$$MSE(\mathcal{M}_T, \mathcal{M}_n) = \|\mathcal{M}_T - E\mathbf{a}\|^2 + \sum_{j=1}^{\mathsf{card}\mathcal{M}} \mathsf{Var}(a_j). \tag{17}$$

In (17), $E\mathbf{a}$ is used as the location in place of the w_js. This can be justified for n large enough provided the estimators a_j are consistent for their means; similar remarks apply to the cases below.

Second, generic model weights may be regarded as a vector $(w_1, ..., w_{\mathsf{card}(\mathcal{M})})$ with one entry for each model in $\mathcal{M}$. So, the model weights for $\mathcal{M}_T$, namely $(w_1^T, ..., w_{m_T}^T)$, can be embedded into the weight vector $\mathbf{w}^T = (w_1^T, ..., w_{\mathsf{card}(\mathcal{M})}^T)$ in terms of $\mathcal{M}$. Each element w_j^T of $\mathbf{w}^T$ is now either the corresponding element in the vector for $\mathcal{M}_T$ or zero. Also, the estimated model weights $\hat{\mathbf{w}}_n$ for $\mathcal{M}_n$ can be embedded into a vector $\hat{\mathbf{w}}$ of length $\mathsf{card}(\mathcal{M})$. Again, $\hat{w}_j$ in $\hat{\mathbf{w}}$ is either the corresponding estimated weight $\hat{w}_j = \hat{w}_j(\mathcal{D})$ or is zero. Now, the variance-bias expression for the model weights is,

$$MSE(\mathbf{w}_T, \hat{\mathbf{w}}) = \|\mathbf{w}_T - E\hat{\mathbf{w}}\|^2 + \sum_{j=1}^{\mathsf{card}\mathcal{M}} \mathsf{Var}(\hat{w}_j). \tag{18}$$

Third, like the model list and the weights, the parameters can also be embedded in a higher order space. Write $\mathbf{B} = (\beta_1, ..., \beta_{\mathsf{card}(\mathcal{M})})$ to mean the collection of parameters

from the models in $\mathcal{M}$ and let $(\beta_1, ..., \beta_{\mathsf{card}(\mathcal{M})})$ be the parameters from the models in $\mathcal{M}_T$. Then we can define $\mathbf{B}_T$ to represent the parameters from the models in $\mathcal{M}_T$; i.e., an entry β_f^T in $\mathbf{B}_T$ is β_f^T if $f \in \mathcal{M}_T$ and is $\beta_f = 0$ for $f \in \mathcal{M} \setminus \mathcal{M}_T$. Similarly, $\hat{\mathbf{B}}$ is the vector of estimated parameters with entries $\hat{\beta}_f$ for $f \in \mathcal{M}_n$ and $\hat{\beta}_f = 0$ for $f \in \mathcal{M} \setminus \mathcal{M}_n$. Now,

$$MSE(\mathbf{B}_T, \hat{\mathbf{B}}) = \sum_{f \in \mathcal{M}} \mathsf{bias}(\hat{\beta}_f, \beta_f^T) + \sum_{f \in \mathcal{M}} \mathsf{Cov}(\hat{\beta}_f). \tag{19}$$

Taken together, the three equations (17), (18) and (19) are unified bias-variance analysis for the stacking predictor.

To see how a unified bias-variance analysis can include inputs that are not estimable, consider a new predictor, the BMA. Given a prior W across the models in $\mathcal{M}$ and a set of priors $\{W_f |\ f \in \mathcal{M}\}$ for the parameters β_f, the BMA can be written

$$\hat{Y}_{n+1} = \sum_{f \in \mathcal{M}} W(f|\mathcal{D}) X_f E_{W_f}(\beta_f|\mathcal{D}). \tag{20}$$

It is well-known that BMA is optimal in an L^2 sense, see Dempster (1973) . Now, the estimable inputs are the model list $\mathcal{M}_n$, the model weights $W(f|\mathcal{D})$, and the parameter estimates, here taken to be posterior means, $E(\beta_f|\mathcal{D})$. The non-estimable inputs are the within model priors and the across model priors. Since it is obvious how to adapt (17), (18) and (19) to the BMA setting, it is enough to give the extra expressions for the variances of the two kinds of priors used to form the BMA. Note that priors do not have a bias because they are not estimable.

To give a variance for a single W_f on β_f recall Gustafson and Clarke (2004) and regard W_f as an element of a class of priors indexed by $\lambda = \lambda_f \in S_f$. That is, there is some $\lambda_0 \in S_f$ so that $W_f = W_{\lambda_0}$. Since λ_f is a hyperparameter, write its distribution as Π_f. Now, by the conditional covariance identity,

$$\mathsf{Cov}(\beta_f|\mathcal{D}) = E_{\Pi_f} \mathsf{Cov}_{W_f}(\beta_f|\mathcal{D}, \Lambda_f) + \mathsf{Cov}_{\Pi_f}(E_{W_f}(\beta_f|\mathcal{D}, \Lambda_f)), \tag{21}$$

where Λ_f is the random variable with distribution Π_f taking values $\lambda_f \in S_f$. The first term in (21) is the usual parameter uncertainty averaged with respect to Π_f across the possible priors, in principle it is already included in the variance expression for the parameters as in (19). The second term, however is new. It is the uncertainty about the estimator $E(\beta_f|\mathcal{D}, \lambda)$ of the parameter from the across-prior variation due to Π_f. This second term therefore represents the variance of the prior, relative to Λ_f.

An analogous technique can be used to give a variance for W. First, regard W as an element of a class of priors $\mathcal{P}$ indexed by $\psi \in S$ with associated random variable Ψ having distribution $\mathcal{Q}$. Let the vector of model weights be

$$\mathbf{w} = (W(1|\mathcal{D}), \ldots, W(\mathsf{card}\mathcal{M}|\mathcal{D})), \tag{22}$$

where $i = 1, ..., \mathsf{card}\mathcal{M}$ indexes the f's in $\mathcal{M}$. Extending (22) to include the hyperparameter ψ leads to

$$\mathbf{w}_\psi = (W_\psi(1|\mathcal{D}), \ldots, W_\psi(\mathsf{card}\mathcal{M}|\mathcal{D})). \tag{23}$$

Now, let us use $\mathbf{w}_\psi$ in place of $\mathbf{w}$ to evaluate variability. Let $\mathsf{Dir}(\mathbf{w}_\psi)$ denote the Dirichlet distribution on the $\mathsf{card}\mathcal{M} - 1$ dimensional simplex of weights $(w_1, ..., w_{\mathsf{card}\mathcal{M}})$ on the models in the BMA. Again, the conditional covariance identity can be applied. It gives

$$\mathsf{Cov}(W|\mathcal{D}) = E_{\mathcal{Q}}\mathsf{Cov}_{\mathsf{Dir}}(\mathbf{w}|\mathcal{D}, \Psi) + \mathsf{Cov}_{\mathcal{Q}}E_{\mathsf{Dir}}(\mathbf{w}|\mathcal{D}, \Psi). \tag{24}$$

The first term in (24), like the first term in (21), represents the uncertainty in the model weights as parameters that is accounted for in (18) and the second term in (24), like the second term in (21) is the uncertainty about the estimator $E_{\mathsf{Dir}}(\mathbf{w}|\mathcal{D}, \psi)$ of the model weights from the across-model prior variation due to $\mathcal{Q}$. Gustafson and Clarke (2004) show how this reasoning can be extended to assign a variance to the choice of space of functions used in models; this would give a third term in (25) below.

Now, the extra variation due to the inputs that cannot be estimated is from (21) and (24), given by

$$\sum_{f\in\mathcal{M}} \mathsf{trace}\left(\mathsf{Cov}(E(\beta_f|\ \mathcal{D}, \Lambda_f)) + \mathsf{Cov}(E(\mathbf{w}|\ \mathcal{D}, Q)\right). \tag{25}$$

So the unified variance bias decomposition for BMA is given by the analogs of (17), (18), (19) and (25).

In some simple cases, an elaborate bias-variance analysis is not explicitly necessary. Consider a binary classification problem with many explanatory variables under zero-one loss. The model list would be very large including some simple models and some very complex models. Let the data be partitioned as $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ with $\mathcal{D}_1 \cap \mathcal{D}_2 = \phi$ to represent training and testing. Suppose a simple model built on $\mathcal{D}_1$ was found to perform quite well on $\mathcal{D}_2$. It might be decided, for instance, that this sort of verification were enough if squared-error CV were used since it seeks it a good tradeoff between bias and variance. That is, under enough assumptions (light tails in the error, the validity of a parsimonious model etc) squared error CV might be deemed de facto equivalent to a bias-variance analysis concluding that all the terms were satisfactorily small. One benefit of a bias-variance analysis is that it may reveal how weak predictors can be improved. A specific example is given in Ransohoff et al. (2008) who tries to correct excessive variance and bias in earlier studies developing molecular markers for colorectal cancer. This paper develops a classifier with a sensitivity of 78 %, a specificity of 53 %, and an accuracy of 63% – better than chance and enough to motivate further work but not really good enough for clinical use.

As a fifth example, suppose we have a data that is a long nonstochastic vector and recall the use of the log-scoring rule at the end of the discussion of Desideratum #1. The nonstochasticity means that variances of "estimators" are not reasonable, however, expressing the typical distance between a prediction and an outcome as a bias may not be unreasonable. Indeed, for the Shtarkov solution (7) in the Bayes setting, it is enough to give a bias-variance bias analysis for (4). Doing so necessitates a concept of the "true" action a density denoted by, say, $q_T(x_i)$. There are several choices for q_T, as will be discussed shortly. Whichever q_T is chosen, difference in data compression properties

between $q_{opt}(\cdot)$ and $q_T(\cdot)$ is

$$\ln \frac{1}{q_T(x^n)} - \ln \frac{1}{q_{opt}(x^n)} = \ln \frac{q_T(x^n)}{q_{opt}(x^n)}. \tag{26}$$

Taking the expectation in the true density gives the relative entropy $D(q_T \| q_{opt}) \geq 0$. Let Γ be a convex set of densities so that $q_T \in \Gamma$. Then, the information projection q_{proj} of q_{opt} satisfies

$$D(q_T \| q_{opt}) = D(q_T \| q_{proj}) + D(q_{proj} \| q_{opt}) \tag{27}$$

where

$$D(q_{proj} || q_{opt}) = \min_{p \in \Gamma} D(p \| q_{opt}),$$

see Csiszar (1975), (eq. 1.5,7). The first term in (27) measures how close the best approximation of q_{opt} within Γ is to the true q_T. The second term in (27) represents how close q_{opt} is to the best approximation of q_T within Γ. In this case, the variance-bias analysis consists of two terms, both of which look like biases. In this case there is no natural analog to a variance; this contrasts with the expansion obtained in Gustafson and Clarke (2004) which has no bias terms.

There are several natural choices for q_T giving different forms for (27). If one of the members of the parametric family is correct, say $q_T = q_{\theta_T}$ and Γ is the whole parametric family then (27) becomes

$$D(q_{\theta_T} \| q_{opt}) = D(q_{\theta_T} \| q_{\theta_{proj}}) + D(q_{\theta_{proj}} \| q_{opt}), \tag{28}$$

where $q_{\theta_{proj}}$ is information projection of q_{opt} onto the parametric family. If θ_T and θ_{proj} are close, the first term in (28) is $(1/2)(\theta_T - \theta_{proj})^2$ by a Taylor expansion, reinforcing its interpretation as a bias. Note that technically, q_{opt} outperforms q_T predictively in the log scoring rule. Nevertheless, if q_T is true, it is reasonable to compare q_{opt} to it.

Another natural choice for q_T is to choose a convex class of priors $\mathcal{W}$ on $\{p_\theta | \theta \in \Theta\}$ containing a baseline prior $w(\theta)$. It is known that asymptotically, the distance between $m(x^n)$ and q_{opt} converges to zero. In essence, this assumes one of the values of θ is correct, but we do not know it. Write $m_\nu(x^n) = \int \nu(\theta) p_\theta(x^n) d\theta$ for the marginal for the data with respect to any $\nu \in \mathcal{W}$. Then, if $q_T = m_w$, the set $\Gamma = \{m_\nu | \nu \in \mathcal{W}\}$ is convex and we have

$$D(m_w \| q_{opt}) = D(m_w \| q_{proj}) + D(q_{proj} \| q_{opt}), \tag{29}$$

where $q_{proj} = m_\nu(x^n)$ given by mixing with respect to $\nu \in \mathcal{W}$ is the information projection of q_{opt} onto the class of all mixture distributions. Now, the first term is

$$D(m_w || q_{proj}) = \int w(\theta') q(x^n | \theta') \ln \frac{m_w(x^n)}{q(x^n | \theta')} d\theta' dx^n \tag{30}$$

$$+ \int w(\theta') q(x^n | \theta') \ln \frac{q(x^n | \theta')}{m_\nu(x^n)} d\theta' dx^n \tag{31}$$

with obvious modifications for the cumulative conditional predictors in (7). It is seen that (1/n) times (30) is asymptotically $(I_\nu(\Theta; X^n) - I_w(\Theta; X^n))/n = \mathcal{O}(1/n)$, see Clarke and Barron (1994), where $I_w(\Theta, X^n)$ is the Shannon mutual information and Θ has distribution w. As n increases, this is $\mathcal{O}((\ln n)/n)$. The second term in (29) is, using (7) with a flat prior,

$$D(q_{proj} \| q_{opt}) = \sum_{i=1}^{n} E_{m_w} D(m_w(x_i|x^{i-1}) || q_{opt}(x_i|x^{i-1})). \tag{32}$$

If this sum is well-behaved, it will be $\mathcal{O}(n)$ and the average will be $\mathcal{O}(1)$. Again, it is reasonable to regard both terms in (29) as biases when m_w is taken as true.

Note that (28) and (29) can be used with any convex Γ and any $q_T \in \Gamma$. It would be most natural to choose q_T and Γ to be in a region of densities thought to be true. Thus, a third choice for q_T is to let Γ be a convex set of densities that may contain the parametric family $p(\cdot|\theta)$ but is much larger than $\{p_\theta | \theta \in \Theta\}$. Then, one can seek, for instance, the minimal complexity density estimator, see Barron and Cover (1990) for a given set of data at the cost of using an estimator rather than a genuinely' true' density. This would model the case that we are relatively confident the 'true' density cannot be constructed from the parametric family $\{p_\theta | \theta \in \Theta\}$ and so must be estimated. Let $L(\cdot)$ be a fixed coding scheme for the elements of Γ and recall that the Shannon codelength for x^n with respect to $q \in \Gamma$ is $\ln(1/q(x^n))$. Now, the two stage codelength for x^n is

$$\ell_\Gamma(x^n, q) = L(q) + \ln \frac{1}{q(x^n)}. \tag{33}$$

Using Γ enlarges the problem so the performance of $q_{opt}(\cdot)$ can be assessed relative to

$$\hat{q}_T(x^n) = \arg \min_{q \in \Gamma} \ell_\Gamma(x^n, q).$$

Again, we have an expression analogous to (28) and (29). However, now, the first term is $D(\hat{q}_T || q_{proj})$ which is data dependent. In the special case that some $\theta \in \Theta$ is true, $\hat{q}_T$ will behave much like $q_{\hat{\theta}}$ where $\hat{\theta}$ is the posterior mean. Since there will be $q_{proj} = q_{\theta_T}$, the first term becomes much like $E(\hat{\theta} - \theta_T)^2$ which has a conventional bias variance decomposition. Thus, we would find that using an estimator such as $\hat{q}_T$ leads to a three term decomposition of a variance term and two bias terms (unless the bias of $\hat{q}_T$ were of concern as well).

Desideratum #4: Every prediction procedure should have an associated complexity with a meaning formally related to the complexity of the data generator.

Intuitively, the complexity of a prediction task i.e., the difficulty of mimicking the output of a DG perhaps by approximating it, and the complexity of predictors are both related to predictive errors. More complex predictors are often better at approximating a wider range of DGs than less complex predictors and so give better predictive performance when the DG is complex or little is known about it. Less complex predictors are often better at giving serviceable approximations to a DG when a simple approximation is good. In both cases, a good approximation means small prediction errors and this is dependent on the complexity of the DG. Note that the converse often holds as well. That is, if the complexity of a predictor is allowed to be too large relative to the approximation task, then ridiculous predictors with seemingly small prediction errors can be found while if the complexity of a predictor is too small then the best approximation may still be quite poor.

The complexities of predictors and prediction problems are important because, under the Prequential principle alone, two predictors giving the same sequence of predictions are indistinguishable even though one must be chosen. It can be seen that bias-variance by itself is not enough because an approximation task does not intrinsically have a meaningful bias or variance although it does have a complexity; this can be measured in a variety of ways including the number of terms required to give an approximation of prespecified exactitude on a domain, the entropy of the DG or its Kolmogorov complexity. The implication is that it is often good strategy to seek a predictor with a complexity appropriate to the prediction problem . The belief undergirding Desideratum #4 is that when a predictor of the appropriate complexity is used, its complexity will match the complexity of the data generator or the intrinsic difficulty of approximating it, see Clarke and Clarke (2009).

There are numerous ways to define the complexity of a prediction task and of a predictor. The most obvious possibility is based on a unified bias-variance analysis. The idea is to generalize from the bias-variance decomposition in (8) by merely defining the complexity of the predictor to be the sum of its squared bias and variance terms. Then, the minimum of the sum can be taken as the definition of the complexity of the prediction task.

In the simplest case, the prediction problem devolves to estimating a real parameter and the minimal MSE can be regarded as a complexity of approximating the true model. Separately, a predictor has a complexity and the most common way to express this is by its bias and variance. Then, the usual bias-variance decomposition (8) is the relationship between the complexity of the prediction problem and the complexity of a predictor. When the minimal MSE is achieved by a predictor, the predictor has optimal complexity in the MSE sense. More generally, Desideratum #4 requires an evaluation of how complex or difficult the prediction problem is, how complex the predictor is (expressed perhaps as a sum of difficulty terms involving biases and variances), and a relationship between them.

To see this explicitly, observe that the unified bias-variance analyses for Desideratum # 3 leads to two expressions for the complexity of the prediction problem, one for each choice of predictor, namely, stacking and BMA. For the stacking predictor, the unified variance-bias analysis leads to a complexity given by the sum of (17), (18), and (19). So, the complexity of the prediction problem i.e, of the function approximation, may be taken as

$$\begin{aligned}\mathcal{C}(\mathsf{stacking}) &= MSE(\mathcal{M}_T, \mathbf{w}_T, \mathbf{B}_T; \mathcal{M}_n, \hat{\mathbf{w}}, \hat{\mathbf{B}}) \\ &= MSE(\mathcal{M}_T, \mathcal{M}_n) + MSE(\mathbf{w}_T, \hat{\mathbf{w}}) + MSE(\mathbf{B}_T, \hat{\mathbf{B}}).\end{aligned}$$

In the case of BMA, we must include the two extra terms from (25). So, the complexity of the prediction task is

$$\begin{aligned}\mathcal{C}(\mathsf{BMA}) &= MSE(\mathcal{M}_T, \mathbf{w}_T, \mathbf{B}_T; \mathcal{M}_n, \hat{\mathbf{w}}, \hat{\mathbf{B}}) + \sum_{f \in \mathcal{M}} \mathsf{trace}\,(\mathsf{Cov}(E(\beta_f |\ \mathcal{D}, \Lambda_f)) \\ &+ \mathsf{Cov}(E(\mathbf{w} |\ \mathcal{D}, Q)),\end{aligned}$$

in which $\mathbf{w}$ corresponds to the Bayes weights not the stacking weights.

Both of these complexity expressions represent the difficulty of the prediction problem in terms of a sum of biases and variances associated to a predictor. However, a single prediction task now has two different complexities. Stacking and BMA can both be used to make predictions for the same problem, but treating the sum of their bias and variance terms as a complexity leads to $\mathcal{C}(\mathsf{stacking})$ and $\mathcal{C}(\mathsf{BMA})$ which are not equal. Thus, the natural choice for the MSE complexity is the minimum over a large class of all predictors that contains BMA and stacking.

Note that treating the sum of a predictor's variances and squared biases as a complexity is a bit of a cop out because because less trivial alternatives are undeveloped. Indeed, in Section 4 an example is given in which the complexity of the predictor is defined by the extent of the search over models that it uses and the complexity of the problem is assessed by how large a class of functions is needed to approximate the DG well in terms of CPE. There it will be seen that Desideratum #4 is consistent with a sort of "Principle of Matching" between the complexity of the approximation and the complexity of the predictor, cf. Clarke and Clarke (2009). The best predictor seems to have a complexity matched to the complexity of the DG. The intuition is that when a predictor is more complex than the DG, the predictor searches through so many functions that its performance can be worse than that of a simpler predictor which does a smaller search. Likewise, when a predictor is simpler than the DG, the predictor cannot accommodate the complexity of the DG so its performance can be worse than a more complex predictor that tracks the DG more readily.

Desideratum #5: Generate a Comprehensive Robustness Analysis.

Robustness asks that a prediction strategy perform well even when the setting is changed. For instance, that a predictor's performance not be affected too much when

the data is bad (outliers, missing data) or when the data generator is different from anything the predictive scheme in intended to model. Thus, varying the true model for which a predictor is found would be a matter of robustness, while comparing different predictors for the same data generator would typically be within a bias-variance analysis.

In contrast to biases and variances which we want as small as possible, robustness should be at the "right" level: Too much robustness leads to posterior insensitivity and too little leads to posterior instability. It is not generally obvious how much robustness is optimal. Nevertheless, predictors that have good robustness properties have a greater claim to plausibility than those that don't.

A comprehensive robustness analysis parallels the unified bias-variance analysis in Desiderata # 3 and used in # 4. The idea is that each aspect of the setting in which a predictor scheme is to be used must be varied the right amount to ensure that each output of a predictor scheme is satisfactorily stable. The outputs include the CPE and features of the predictor itself among others. We distinguish two sorts of robustness, local and global, depending on which aspect of the setting is being varied. First, local robustness concerns aspects of the setting that are objective; they are fixed and deterministic, usually arising from modeling assumptions believed to be true. The data are in this category, too. The natural way to evaluate robustness in these cases is through local perturbations: Choose a neighborhood around each of the nominal inputs and ensure that the consequent predictions do not vary overmuch. The local perturbations should capture the small deviations from the model that one would expect the experimental set up to have.

By contrast, global robustness concerns aspects of the setting which are subjective, not justified by any physical understanding of the DG. Thus, it makes sense to model these perturbations as stochastic, invoking the Principle of Insufficient Reason; see Kass and Wasserman (1996). These perturbations are, therefore, not generally local. Compared to the ranges used in local perturbations for deterministic inputs, these are large scale perturbations. In this case, one would be led to choose a hyperprior that was objective in some sense, possibly close to a uniform.

Since this is rather abstract, consider two examples. First, an objective prior W might arise from an optimality principle that represented modeling information believed to be true. As a consequence, W would be the only reasonable prior to use, unless the modeling information were called into question. This implies that varying W locally makes sense as a way to test the stability of the CPE against small perturbations in the efficacy of the assumed information. By contrast, if no objective information went into the choice of W it would make sense to exhibit W as say W_{λ_0}, a prior in a class defined by the hyperparameter λ to which a distribution $\pi(\lambda)$ was assigned. This would be one way to test W by mimicking its subjective selection as a random process. That is, it might make sense to regard $W = W_{\lambda_0}$ as an outcome of Λ and so consider the marginal distribution from integrating over λ. The distribution assigned to Λ would be as dispersed as reasonably possible and again the effect on CPE would be evaluated.

As a second example, consider the model space in a regression problem. If it were believed that the function really was a waveform then a Fourier analysis would be

appropriate for physical reasons. In this case, it would make sense to evaluate the effect on CPE of varying the Fourier basis locally. For instance, one could vary the Fourier basis by replacing $\sin(nx)$ with $\sin(nx + \delta(x))$ for some small function δ or $\sin^{1+\alpha}(nx)$. These would be quite difficult, so it would probably be enough for evaluating local robustness to vary the Fourier coefficients over a small neighborhood as a proxy for varying the Fourier basis. By contrast, if there were no extra modeling assumptions that were reasonable, a large scale perturbation or global robustness would be appropriate. For instance, re-estimating the regression function using Legendre polynomials would assess the robustness of the predictions to the choice of basis.

Just as there is a desire that all the biases and variances be combined in a single notion of problem difficulty, so it is important to consider the relative contributions of the different components of robustness to an overall robustness. To see this, suppose there are two aspects of the setting in which a predictor is to be used, say a_1, a_2 and one output o, CPE for instance, with baseline inputs $\tilde{a}_i$, $\tilde{a}_2$ and corresponding baseline output $\tilde{o}$. Then an overall sensitivity can be evaluated as follows. Choose distances d_j on i_j, $j = 1, 2$ and then set

$$d_i((a_1, a_2), (\tilde{a}_1, \tilde{a}_2)) = d_1(a_1, \tilde{a}_1) + d_2(a_2, \tilde{a}_2). \tag{34}$$

Also, choose a distance on the output to be $d_o(o, \tilde{o})$. Solving the optimization problem

$$\max \quad d_o(o, \tilde{o}) \quad \text{subject to} \quad d_i((a_1, a_2), (\tilde{a}_1, \tilde{a}_2)) < \epsilon, \tag{35}$$

for $\epsilon \to 0^+$ can be done through second order matrix approximations in many cases. If there are three or more aspects, then there are three or more terms in (34), respectively. Clarke and Gustafson (1998) use the relative entropy to evaluate the overall sensitivity of the posterior to three aspects, namely the prior, likelihood, and data. The result is a vector corresponding to the maximal eigenvalue indicating the direction among the collection of inputs along which the deviation of o from the baseline $\tilde{o}$ is fastest. The consequence of this is that the relative influence of each input to the output can be evaluated.

Desideratum #6: Determine the limiting properties of the predictors, any estimates they use, and any functions of them that are of interest and verify they do not contradict the context of the problem.

The content of Desideratum 6 is a logical consistency requirement: You don't want to be able to derive something under reasonable conditions that is different from what would be observed. It is this aspect of Coordinating Theory where conventional statistical modeling has the most to offer. It is commonplace to evaluate the limiting properties for classifiers, estimators, and other ingredients that form part of a prediction scheme, under the concept of a true model which does not figure in the earlier Desiderata. Evaluating predictors is helpful for revealing their properties. However, for many real data sets where the concept of a true model is problematic the behavior of a predictor under a candidate true model is only in a contingent sense. So, the best we can expect is to learn when predictors have theoretical properties that overtly contradict the empirical performance we want.

4 The Desiderata and Other Theories

In this section, we examine the interplay between Coordinating Theory, Bayes theory and Frequentist theory. The three approaches have some common features, but Coordinating Theory is closer to Bayes.

Before going into these details, note that the desiderata are mostly a formalized way to evaluate a predictor class. That is, one uses the desiderata to determine systematically how good a predictor class is for a given DG. To apply the desiderata, the predictor must be exhibited as an output of its inputs and then each desideratum considered in turn. For instance, if the predictor for stochastic data is the predictive density $m(x_{n+1}|x^n)$, the inputs are the prior, the likelihood, and the data in which the prior is subjective and the likelihood is given by modeling assumptions. Then, looking at predictive performance and updating (here trivially since the inputs are fixed) satisfies #1 and #2. None of the inputs are estimable, however, a variance can be assigned to the prior as in #3 and a bias can be assigned to the likelihood, possibly as in the non-stochastic case. (If the likelihood were not based on modeling assumptions, then it would be assigned a variance possibly using the conditional covariance approach of situating it in a larger family of likelihoods.) The sum of these terms measures the complexity of the prediction problem as in #4. A local perturbation of the likelihood and data would be combined with a stochastic perturbation of the prior with respect to a near uniform hyperprior and the three would be combined to give a comprehensive robustness analysis of CPE or of the predictor itself for #5. Examining the asymptotics would satisfy #6. Notice that the parameter does not appear explicitly on this list. However, if the predictor were formed using an estimate of θ, say $p_{\hat{\theta}}$, then the variance and bias of $\hat{\theta}$ would be included under desideratum #2.

4.1 Coordinating Theory and Bayes

The Bayesian approach can be regarded as one rational way to construct a predictor. The construction rests on a collection of axioms such as those in Bernardo and Smith (1994) or on Savages postulates, see Savage (1954). The independence of the prior from the data is optimal, see Freedman and Purves (1969). An important sense in which Bayes methods are best is provided by the Complete Class Theorem which states (roughly) that the procedures which are Bayes for a given problem form a complete class, i.e., a class that contains all admissible procedures, see Brown (1981). Taken together, this provides a unified framework for data analysis, hypothesis testing, and prediction.

However, the Bayesian framework makes a simplifying assumption that is not reasonable in general. Once the inputs – loss function, prior, likelihood, etc. – have been specified, they are not updated. It is straightforward, in principle, to use the Bayes approach to get solutions but it is not straightforward to reformulate the Bayes problem e.g., to rechoose the inputs such as a model list or loss function in response to lack of fit or poor prediction.

Despite this, Bayes prediction is asymptotically as good as optimal prediction in many cases. Indeed, Dawid and Skouras (1998) and Dawid and Skouras (1999) establish Bayesian prediction systems are efficient asymptotically and Wong and Clarke (2004) found that optimizing a conditional risk, chosen by using the accumulated data, to predict a future outcome outperformed standard Bayes techniques in small sample sizes, asymptoting to the performance of Bayesian predictors. In the presence of bias, Clarke (2003) has shown that prediction based on stacking can predictively outperform BMA.

Essentially, Coordinating Theory enlarges the Bayes predictive problem with prequential thinking by replacing the notion of credible set with predictive accuracy and invoking Desiderata #2 and #3 to structure it. That is, periodic re-selection of the model list or other elements and going outside the Bayes framework to include bias (and variances) permits the Bayesian to be a better Bayesian by finding the right decision problem to solve and evaluating how well the solution performs.

To see how the desiderata can be applied to BMA, for instance, list the inputs to the predictive procedure. Suppose sequential outcomes Y_i from $Y = F(\mathbf{X}) + \epsilon$ with F unknown but an IID error structure are available where $\mathbf{X}$ is a collection of p explanatory variables. Then, for time step n, we have a collection S_n of models. The models in S_n may have parameters and be written $F_j(\mathbf{X}|\theta_j)$. To form the average we use a prior $w_n(j)$ on the models in S_n, which is updated to form $w(j|\mathcal{D}_n)$, where $\mathcal{D} = \{(\mathbf{X_i}, Y_i) : \quad i \leq n\}$ and equip parameter θ_i with prior $w(\theta_i)$. So, there are six inputs: (1) the θ_j's within the models F_j, (2) the priors on the θ_j's within each F_j, (3) the model lists for each n, (4) the error term, (5) the prior on the models, and (6) the space S_n is drawn from. Note that including the individual models F_j on this list would be redundant from a bias-variance standpoint because varying the prior (5) on the models on the list effectively varies the models conditional on the list while varying the model list (6) necessitates varying the F_j's. From the sensitivity standpoint, local variation of the models would be reasonable only if the models on the list were regarded as true in some sense, as in the Fourier basis earlier and then there would be no point in varying the model list stochastically for bias-variance.

Now, we go through the desiderata in turn. Desideratum # 1 is satisfied if the MSPE satisfies the prequential principle. Desideratum #2 is satisfied if the model lists S_n are reselected as a function of the predictuals; see the discussion preceding Desideratum #3. For instance, one may start with linear functions of the variables in X, find that there are patterns in the residuals, and be led to expand the collection of regression functions to include square and rectangular terms in the X_i's. Doing this may require changing the prior on S_k and assigning priors to the parameters in the extra models included in S_{k+1}.

Desideratum #3 requires variances and biases. The bias-variance decomposition has already been given in Section 2 and is the sum of (17), (18), (19), and (25), the first three suitably modified. Note that of the six overall inputs, the only two left out are the error term and the model space.

Desideratum #4 requires looking at the sum of the terms in the bias-variance decomposition to verify that BMA had effectively minimized them. This would necessitate

calculating all the terms; this can be done, but is the subject of a future paper. However, some heuristics can be easily given. First, if the true model is linear and is on the list S_n and $\dim(x)$ is not large, then the complexity of the prediction task is probably small because the true model is easily approximated, even identified. This is the $\mathcal{M}$-closed case of Bernardo and Smith (1994). Accordingly, in this case, the biases and variances should all be small. Second, in the $\mathcal{M}$-complete case the complexity is probably in some mid-range under any reasonable calibration because the true model is not in any S_n on account of it being so difficult to approximate. So, the biases should be small at the cost of higher variances or the variances should be small at the cost of higher biases. In the $\mathcal{M}$-open case, the situation is more difficult because the true model is not in S_n and is not really approximable by the available models. This last would be a higher bias and higher variance situation. It is in these contexts that a Principle of Matching may be reasonable: For the $\mathcal{M}$-closed case low complexity predictors would likely be best; for the $\mathcal{M}$-complete case medium complexity predictors would likely be best, and, for the $\mathcal{M}$-open case, high complexity predictors would likely be best.

Desideratum #5 would apply to the data, the error term and the model space itself, the latter being summarized by a prior over different representations of a model space in terms of basis elements, for instance. In some cases, the formulation in (34) and (35) is amenable to standard quadratic maximization problems, as used in Clarke and Gustafson (1998). Essentially, a quadratic approximation can be given for each term locally so that maximizing a quadratic objective function subject to quadratic constraints becomes feasible. Note that the quantities subject to the sensitivity analysis are not subject to a bias-variance analysis and conversely, but each input to the predictor is examined under Desideratum #3 or #5. There is no prohibition on looking at the sensitivity of CPE, say, to the prior W over models. However, in practice, it may be enough to ensure that $\mathrm{var}(CPE)$ is not too small or too large as a function of W.

Finally, Desideratum #6 means that once we have formalized our procedure, we want to be sure that its limiting properties are reasonable. For BMA, we want to be sure that if there is a true model in S that in the limit S_n converges to a list that includes the true model and that the weight in the BMA on the true model converges to one. If the true model is not in S, then we want to know that the BMA converges to the model in S closest to the true model. Moreover, we expect that the estimates of the parameters in the true model will converge with the usual asymptotic normality, in regular cases.

4.2 Coordinating Theory and Frequentism

The Frequentist approach is another way to construct a predictor. Unlike Bayes, Frequentism does not have a comprehensive foundational construction. The analog to Savage's postulates is roughly provided by the von Neuman-Morgenstern expected utility theorem. However, there is little justification for the use of a sampling distribution for estimation apart from the repeated sampling interpretation from probability theory.

Coordinating Theory replaces the concept of confidence with predictive accuracy

and only uses Frequentist concepts based on the sampling distribution for aggregate characterizations as in Desideratum # 3, with variances taken over the sample space or in Desideratum #6 where asymptotics are used as a sort of sanity check. One could sidestep Frequentism further by using posterior variances in Desiderata #3 and # 4. However, the Frequentist variance is a broader summary of variability since it involves integration over the sample space. Moreover, Coordinating Theory allows Frequentist thinking between time steps in the respecification of the decision problem leading to a predictor, in Desideratum #2. The reselection of a decision problem is non-Bayesian because it rests on correcting bias, a problem that is more readily examined within the Frequentist paradigm than in the Bayesian.

For contrast with BMA, consider the stacking model average as in (14) and (15). Instead of the six inputs, there are now only five since the two sorts of priors are not used but model weights are. The five inputs are (1) the θ_j's within the models F_j, (2) the model weights, (3) the model lists for each n, (4) the error term, and (5) the space S_n is drawn from.

As in the BMA case, we go through the desiderata for the stacking model average predictor. Desideratum # 1 is satisfied if the MSPE satisfies the prequential principle. Desideratum #2 is satisfied if the model lists S_n are reselected as a function of the predictuals. The bias-variance decomposition required for Desideratum #3 has already been given in Section 2 and is the sum of (17), (18), and (19). This gives a version of the complexity for Desideratum #4. Desideratum # 5 asks for a robustness analysis. This would proceed as in the BMA case since there are three inputs that have not been assigned biases or variances, the model space and the error term, and the data that must be assessed. Finally, Desideratum #6 would be examined much as in the BMA case but from a Frequentist standpoint.

5 Aspects of Desiderata 1, 2, and 4

In this section, an instance of the kind of analysis suggested by the desiderata is presented. CPE error is calculated, and #1 and # 2 are satisfied. Although neither a bias-variance analysis nor a sensitivity analysis is done, some heuristics on complexity can be given because nine different predictors based on three different function classes and three different model averaging strategies are used.

To begin, consider three model spaces: linear models (LMs), generalized additive models (GAMs), and recursive partitioning models (trees). Also, consider three model average prediction schemes: BMA, here called likelihood weighted averaging (LWA) because of the model list re-selection, stacking, and a data dependent convex combination of these here called adaptive convex average of predictors (ACAP). Because of the relative richness of the function classes they represent, it is reasonable to regard LM, GAM and trees as low, medium, and high complexity respectively. Because of the scope of the search over function spaces and model weights that they involve, it is also reasonable to regard LWA, stacking and ACAP as low, medium, and high complexity, respectively.

To specify the predictors, form the LWA for LMs as

$$\hat{Y}_{LWA,i+1} \sum_{j \in \mathcal{M}_{i+1}} W(j|Data) X_j \hat{\beta}_j \tag{36}$$

and the stacking average for LMs as

$$\hat{Y}_{CV,i+1} = \sum_{j \in \mathcal{M}_{i+1}} \hat{w}_{i,j} X_j \tilde{\beta}_j, \tag{37}$$

where $Data$ includes the first i data points, X_j means the values at time $i+1$ of the jth model, and the $\hat{w}_{i,j}$'s are the stacking weights (at time i) parallel to the Bayes weights in LWA (based on a uniform prior). In (37), 5-fold CV was used. The difference between (36) and (37) is seen in the weights as well as the parameter estimates because (36) used ordinary least squares estimators $\hat{\beta}_j$ and (37) used posterior means $\tilde{\beta}_j$. Now, in principle, it will be possible to compute the MSPE along a string of data so that Desideratum #1 is satisfied.

Note that $\hat{Y}_{LWA,i+1}$ and $\hat{Y}_{CV,i+1}$ may be formed from different model lists denoted $\mathcal{M}_{LWA,i+1}$ and $\mathcal{M}_{CV,i+1}$. We expect these two model lists to be different because stacking and LWA have different properties. Stacking generally has a higher variability than LWA and has slightly better performance under bias. So, when the true model is far from the models on the list stacking often does better than LWA but when the true model is on the model list LWA usually does better than stacking. Thus, the degree of model mis-specification largely determines which method will do better predictively.

For comparison, consider an adaptive combination of an average of the LWA and stacking predictors, ACAPs. That is, for $\hat{Y}_{LWA,i+1}$, $\hat{Y}_{CV,i+1}$, and model lists for each, take another mixture. Let $\alpha_i \in [0,1]$ and define a convex combination of $\hat{Y}_{LWA,i+1}$ and $\hat{Y}_{CV,i+1}$ from (36) and (37) to be

$$\hat{Y}_{i+1,ACAP} = \alpha_i \hat{Y}_{CV,i+1} + (1-\alpha_i)\hat{Y}_{LWA,i+1}, \tag{38}$$

where, for each i, $\hat{\alpha}_i$ achieves

$$\min_{\alpha} \sum_{u=i-20}^{i} \left(Y_u - [\alpha \hat{Y}_{CV,u} + (1-\alpha)\hat{Y}_{LWA,u}] \right)^2. \tag{39}$$

Note that (39) only uses most recent 20 data points so α_t won't converge too fast relative to $\mathcal{M}_k$. When $\alpha_i < 0$, we set it to be 0 and when $\alpha_i > 1$, we set it to be 1.

An analogous procedure was used for GAMs and trees. For these two cases, however, a selection of variables is treated as a model. Given a selection of variables, a GAM can be formed using them and given several GAMs of this form a stacking average found. For the LWA case, several GAMs are formed using different selections of variables and the Bayes weights for them are found using the deviance criterion in the contributed R package gamlss. Likewise for trees, a stacking average is found using several selections

of variables and the LWA is found using the deviance criterion in the contributed R package tree. In both cases, the ACAP average is found using (38).

It remains to specify how the model lists update in response to CPE so that Desideratum #2 will be satisfied. The method is described fully in Clarke and Clarke (2009). For LMs, the basic idea is as follows. Use an ensemble of terms which consists of all terms of second order or less, i.e., all x_j's and all $x_j x_k$'s. A random selection of these terms is used for the initial model lists $\mathcal{M}_{LWA,0}$ and $\mathcal{M}_{CV,0}$ of fixed cardinality $K = 4$. The next step is to define a central model within $\mathcal{M}_{LWA,0}$ and $\mathcal{M}_{CV,0}$. The central model for $\mathcal{M}_{LWA,0}$ contains the terms that are in a majority of the models in $\mathcal{M}_{LWA,0}$ and the central model for $\mathcal{M}_{CV,0}$ contains the terms that are in a majority of the models in $\mathcal{M}_{CV,0}$. For both averaging procedures, new models formed by adding or deleting one term from the central model in all possible ways are considered, and the best of these is compared to the models already on the list. The new model replaces one of the old models if it gives better CV performance. For GAM's this procedure is modified by building models using cubic splines based on the terms in the ensemble. For trees this is modified by looking only at the terms in the selection of variables fed into the tree software, not at the terms in the actual model output by the software.

To compare the 9 methods (3 predictors, 3 model spaces) we used the benchmark data set *Comp-Activ* containing records of computer performance measures used to predict the fraction of time that CPUs run in user mode. This data set is considered reasonably complex. Here, we chose only 15 of the original 24 predictors. The CPE's of the three methods for each model space are in Figs. 1, 2, and 3.

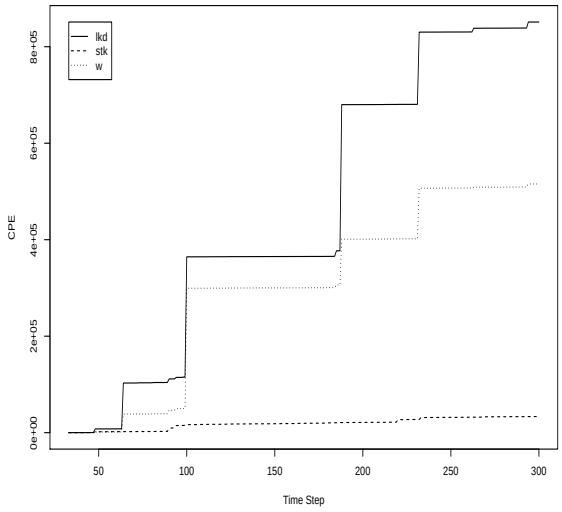

Figure 1: CPEs for LMs $n = 300$ for the Comp-Active Data. The results from fifteen random permutations of the data were averaged.

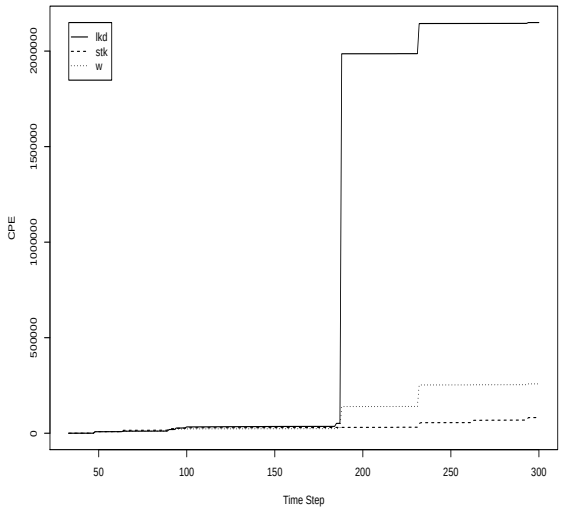

Figure 2: CPEs for GAMs, $n = 300$, for the Comp-Active Data. The results from fifteen random permutations of the data were averaged.

Note that the scales on the vertical axis are different for the 3 graphs. The first is the largest (e^{05} means 100,000), the third is the smallest. All 3 of the smallest final CPEs occur for tree models and among these the ACAP error was less than the LWA error which was less than the stacking error. For LM's and GAM's, the stacking error was less than the ACAP error which was less than the LWA error. The worst were ACAP's and LWA for LM's.

We interpret this to be broadly consistent with a sort of Principle of Matching: The complexity of the data i.e., the approximation task represented by the DG, should match the complexity of the predictor for best performance. In this case, the data set is quite complex and the complexity of the model space is the most important feature of the predictor. So, the best cases occurred for the most complex class, namely trees. The second most complex class was GAMs and it did second best while LMs, the simplest class, gave the worst performance.

Within the most complex model space, trees, the most complex method, ACAPs does best, LWA did less well and stacking did worst. The Principle of Matching would have predicted that stacking would do better than LWA because stacking is more flexible than LWA since its weights are not based on a likelihood. However this was not observed. It may be that the richness of the model space permits the higher efficiency of Bayesian methods to outperform the less efficient stacking average. Or, it may just be noise.

The worst cases were with LM's, the simplest models. Among these, stacking did best. One would have expected from the Principle of Matching that the most complex

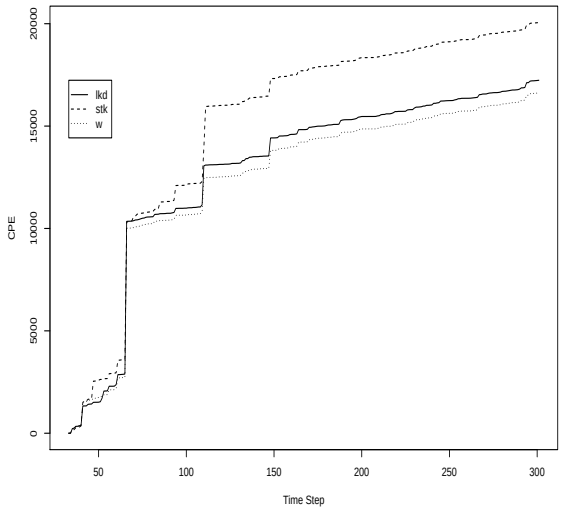

Figure 3: CPEs for Trees, $n = 300$, for the Comp-Active Data. The results from fifteen random permutations of the data were averaged.

predictor ACAP should have done best. However, it is possible that LWA, which converges quickly, has a high bias, making its predictive error so bad it forces the ACAP to do poorly.

6 Implications of the Theory

The point of this paper is to argue the merit and feasibility of expanding the statistical problem so it is phrased in terms of prediction. The 6 proposed desiderata are merely a way to structure thinking about such a general problem. It is worth stating several immediate conclusions that are implied by adopting a predictive standpoint.

First, physical interpretation of models is downweighted in favor of actual predictive performance. Recall that, in practice, many statisticians use physical interpretations to choose among models or predictors. This is often done by preferring a model with a good physical interpretation over one with a poor physical interpretation – even when the latter performs better predictively. That is, physical interpretability is taken as a proxy for validation. Strict application of the Prequential principle will often rule out this use of physical interpretations of models on the grounds that they rest on suppositions about the true model. The proper place for physical interpretations is post-processing an optimal predictor.

Second, for the Bayesian, there is no prohibition on data dependent priors. Wasserman (2000) uses data dependent priors, and shows they are optimal, in a mixture setting

where model uncertainty is a problem. Clarke and Yuan (2004) give general information theoretic interpretations and the sequential procedure of Wong and Clarke (2004) can be interpreted as using a prior that depends on the data.

Third, the posterior variances commonly reported are conditional on the model, and much else, being correct. As soon as model uncertainty is included via updating predictors for instance, the predictive analog of confidence or credibility bands becomes much larger, see Draper (1995). It may well be that once proper modeling uncertainty is taken into account the confidence or credibility bounds become so large as to reveal that the inferences based on the data are merely suggestive. This is another way to say that too many of our inferences have been based on overfitting.

Fourth, the stability or robustness of models is as important as any claim to veracity. Getting a prediction scheme whose predictive error doesn't change overmuch under reasonable perturbations of its inputs is hard enough; insisting that the real world conform to models we can write down conveniently may be unrealistic. Indeed, the best predictors do not usually correspond to any obvious model for the DG. It may be important to back off from model identification in order to derive modeling schemes that give predictions with assessable reliability.

Fifth, and most important, validation of models in the sense of doing extensive checks that the predictions they give are accurate is the most central property of any scheme. Sequential prediction is merely one natural way to do this. Pragmatically, trying to find a model that can be taken as true by surmising validity from weak validation criteria such as interpretability will often lead to overstatement of the strength of the information in complex data sets. Otherwise put, all too often the result will be models that fit adequately but fail to generalize. If a "true model" must be announced, the best way may not be to seek a model directly but instead to develop a good predictor. Then, the predictor can be used to identify a model by converting candidate true models into predictors and finding one that is not too far from the established good predictor.

References

Barron, A. (1993). "Universal approximation bounds for superpositions of a sigmoidal function." *Trans. Inform. Theory*, 39: 930–944. 286

Barron, A. and Cover, T. (1990). "Minimum complexity density estimation." *Trans. Inform. Theory*, 37: 1034–1054. 301

Bernardo, J. and Smith, A. F. M. (1994). *Bayesian Theory.*. Chichester: J. Wiley and Sons, 1 edition. 306, 308

Brown, L. (1981). "A complete class theorem for statistical problems with finite sample spaces." *Ann. Statist.*, 9: 1289–1300. 306

Chipman, E., H. George and McCulloch, R. (1992). "The practical implementation of model selection." In *Model Selection, Lahiri, P., Ed.*, volume 38, 113–126. Beachwood, OH: IMS Lecture Notes-Monograph Series. 285

Clarke, B. (2003). "Comparing Bayes model averaging and stacking when model approximation error cannot be ignored." *J. Mach. Learning. Res.*, 4: 683–712. 288, 307

— (2007). "Information optimality in Bayesian models." *J. Econ.*, 138: 405–429. 292

Clarke, B. and Barron, A. (1994). "Jeffreys' prior is asymptotically least favorable under entropy risk." *J. Stat. Planning Inference*, 43: 37–60. 301

Clarke, B. and Gustafson, P. (1998). "On the sensitivity of the posterior distribution to its inputs." *J. Stat. Planning Inference*, 71: 137–150. 305, 308

Clarke, B. and Yuan, A. (2004). "Partial information reference priors." *J. Stat. Planning Inference*, 123: 313–345. 314

Clarke, J. and Clarke, B. (2009). "Prequential analysis of complex data with adaptive combined average predictors." *Stat. Anal. Data Mining*, 2: 274–290. 302, 303, 311

Csiszar, I. (1975). "Partial information reference priors." *Ann. Probab.*, 3: 146–158. 300

Dawid, A. P. (1982). "The well-calibrated Bayesian." *J. Amer. Stat. Assoc.*, 77: 605–610. 288

— (1984). "Statistical theory: The prequential approach." *J. Roy. Stat. Soc. Ser. B*, 147: 278–292. 283, 287

— (1992). "Prequential data analysis." In *Current Issues in Statistical Inference: Essays in Honor of D. Basu, Ghosh, M. and Pathak, P., Eds.*, volume 17, 113–126. Beachwood, OH: IMS Lecture Notes-Monograph Series. 293

— (2004). "Probability, causality, and the empirical world: A Bayes-de Finetti-Popper-Borel synthesis." *Stat. Sci.*, 19: 44–57. 288

Dawid, A. P. and Skouras, C. (1998). "On efficient point prediction systems." *J. Roy. Stat. Soc. Ser. B*, 60: 765–780. 307

— (1999). "On efficient probability forecasting systems." *Biometrika*, 86: 765–784. 307

Dawid, A. P. and Vovk, V. (1999). "Prequential probability: Principles and Properties." *Bernoiulli*, 5: 125–162. 287

de Finetti, B. (1937). "La prévision : ses lois logiques, ses sources subjectives." *Annales de l'institut Henri Poincaré*, 7: 1–68. 287, 288

Dempster, P. (1973). "Alternatives to least squares in multiple regression." In *Multivariate Statistical inference, Kabe, D. and Gupta, R. Eds.*, 25–40. New York: Elsevier. 298

Domingos, P. (2000). "A unified variance-bias decomposition for zero-one and squared loss." In *Proceedings of the Seventeenth National Conference on Artificial Intelligence.*, 564–569. Austin, TX: AAAI Press. 295

Draper, D. (1995). "Assessment and propagation of model uncertainty." *J. Roy. Stat. Soc. Ser. B*, 57: 45–97. 287, 314

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *J. Amer. Stat. Assoc.*, 96: 1348–1360. 286

Fan, J. and Lv, J. (2008). "Sure independence screening for ultrahigh dimensional feature space." *J. Roy. Stat. Soc. Ser. B*, 70: 849–911. 286

Freedman, D. and Purves, R. (1969). "Bayes method for bookies." *Ann. Math. Stat.*, 40: 1177–1186. 306

George, E. (2001). "Dilution priors for model uncertainty." Slides, http://www.msri.org/publications/ln/msri/2001/nle/george/1/banner/01.html. 285

Gustafson, P. and Clarke, B. (2004). "Decomposing posterior variance." *J. Stat. Planning Inference*, 119: 311–327. 298, 299, 300

Haussler, D. and Barron, A. R. (1992). "How well do Bayes methods work for online prediction of ± 1 values?" In *Proceedings of the Third NEC Symposium on Computation and Cognition*, 74–100. Philadelphia, PA: SIAM Press. 291

Haussler, D., Kivinen, J., and Warmuth, M. (1998). "Sequential prediction of individual sequences under general loss functions." *IEEE Trans. Inform. Theory*, 44: 1906–1924. 291

Heskes, T. (1998). "Bias-variance decompositions for likelihood based functions." *Neural. Comp.*, 10: 1425–1433. 295

Huang, J. and Xie, H. (2007). "Asymptotic oracle properties of SCAD-penalized least squares estimators." In *Asymptotics: Particles, Processes and Inverse Problems.*, volume 55, 149–166. Beachwood, OH: IMS Lecture NotesMonograph Series. 286

James, G. and Hastie, T. (1997). "Generalizations of the bias-variance decomposition for predictive error." Technical report, Statsitics Dept., Stanford University. 295

Kass, R. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *J. Amer. Stat. Assoc.*, 91: 1343–1370. 304

Kim, S., Dahl, D., and Vanucci, M. (2009). "Spiked Dirichlet processes prior to Bayesian multiple hypothesis testing in random effects models." *Bayesian Analysis*, 4: 707–732. 286

Kyburg, H. and Smokler, H. (1980). *Studies in Subjective Probability*. NY: John Wiley and Sons. 287

Leeb, H. and Pötscher, B. (2001). "Sparse estimators and the oracle property of the return of Hodges' estimator?" *J. Econometrics*, 142: 201–211. 286

Lijoi, A. and Prünster (2009). "Distribution properties of means of random variables." *Statistical Surveys*, 3: 47–95. 286

Pericchi, L. (2005). "Model selection and hypothesis testing based on objective probabilities and Bayes factors." In *Handbook of Statistics, Bayesian Thinking, Modeling, and Computation, Dey, D. K. and Rao, C. R. Eds.*, volume 25, 115–149. Maryland Heights, MO: Elsevier. 285

Ransohoff, D. (2004). "Rules of evidence for cancer molecular marker discovery and validation." *Nat. Rev. Cancer*, 4: 309–314. 286, 289

— (2005). "Bias as a threat to the validity of molecular marker research." *Nat. Rev. Cancer*, 5: 142–149. 286, 289

Ransohoff, D., Martin, C., Wiggins, W., Hitt, B., Keku, T., Galanko, J., and Sandler, R. (2008). "Assessment of serum proteomics to detect large colon adenomas." *Cancer Epi. Biomarkers and Prev.*, 17: 2188–2193. 299

Rissanen, J. (1996). "Fisher information and stochastic complexity." *IEEE Trans. Inform. Theory*, 42: 40–47. 292

Savage, L. J. (1954). *The Foundations of Statistics*. NY: John Wiley and Sons. 306

Shtarkov, Y. (1988). "Universal sequential coding of single messages." *Problems of Information Transmission*, 23: 3–17. 291, 292

Storlie, C., Bondell, H., Reich, B., and Zhang, H. (To appear). "Surface estimation, variable selection, and the nonparametric oracle property." *Stat. Sinica*. 286

van Erven, T., Grünwald, P., and de Rooij, S. (2008). "Catching up faster by switching sooner: A prequential solution to the AIC-BIC dilemma." Technical report, arXiv:0807.1005v1. 287, 288, 293, 294

Wainwright, M. (2006). "Estimating the wrong graphical model: Benefits in the computation-limited setting." *J. Mach. Learn. Res.*, 7: 1829–1859. 285

Wasserman, L. (2000). "Asymptotic inference for mixture models using data dependent priors." *J. Roy. Stat. Soc. Ser. B*, 62, Pt. 1: 159–180. 313

Webster, J., Gibbs, J., Clarke, J., Ray, M., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., McCorquodale, D., Cuello, C., Leung, D., Bryden, L., Nath, P., Zisman, V., Joshipura, K., Huentelman, M., Hu-Lince, D., Coon, K., Craig, D., Pearson, J., NACC-Neuropathology Group, Heward, C., Reiman, E., Stephan, D., Hardy, J., and Myers, A. (2009). "Genetic control of human brain transcript expression in Alzheimer disease." *Amer. J. Hum. Genetics*, 445–458. 286

Wolpert, D. (1992). "Stacked generalizations." *Neural Networks*, 5: 241–259. 296

Wong, H. and Clarke, B. (2004). "Improvement over Bayes prediction in small samples in the presence of model uncertainty ." *Can. J. Statist.*, 32: 269–283. 288, 293, 307, 314

Xie, Q. and Barron, A. (2000). "Asymptotic minimax regret for data compression, gambling, and prediction." *IEEE Trans. Inform. Theory*, 46: 431–445. 292

Yang, Z. (1997). "How often do wrong models produce better phylogenies?" *Mol. Biol. Evol.*, 14: 105–108. 285

Zhao, Y. and Atkeson, C. (1993). "Some approximation properties of projection pursuit learning networks." In *Advances in Neural Information Processing Systems 4, Moody, J., Hanson, S. and Lippman, R. Eds.*, 936–943. San Mateo, CA: Morgan-Kauffman. 286

Zou, H. (2006). "The adaptive LASSO and its oracle properties." *J. Amer. Stat. Assoc.*, 101: 1418–1429. 286

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *J. Roy. Stat. Soc. Ser. B*, 67: 301–320. 286

Zou, H. and Zhang, H. (2009). "On the adaptive elastic net with a diverging number of parameters." *Ann. Statist.*, 37: 1733–1751. 286

Acknowledgments

The author thanks Jennifer Clarke for help with the computations and figures in Sec. 5. The author also gratefully acknowledges the insight and generosity of two anonymous referees, the Associate Editor, the Editor, and the Editor-in-Chief. Their combined efforts massively improved this paper.